Geom-SPIDER-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization

Gersende Fort, Eric Moulines, Hoi-To Wai

GEOM-SPIDER-EM: FASTER VARIANCE REDUCED STOCHASTIC EXPECTATION MAXIMIZATION FOR NONCONVEX FINITE-SUM OPTIMIZATION

Gersende Fort[⋆] *Eric Moulines*[†] *Hoi-To Wai*[◇]

[⋆] Institut de Mathématiques de Toulouse, Université de Toulouse; CNRS UPS, F-31062 Toulouse Cedex, France
[†] Centre de Mathématiques Appliquées; Ecole Polytechnique; 91128 Palaiseau Cedex, France
[◇] Department of SEEM; The Chinese University of Hong Kong; Shatin, Hong Kong

ABSTRACT

The Expectation Maximization (EM) algorithm is a key reference for inference in latent variable models; unfortunately, its computational cost is prohibitive in the large scale learning setting. In this paper, we propose an extension of the Stochastic Path-Integrated Differential EstimatoR EM (SPIDER-EM) and derive complexity bounds for this novel algorithm, designed to solve smooth nonconvex finite-sum optimization problems. We show that it reaches the same state of the art complexity bounds as SPIDER-EM; and provide conditions for a linear rate of convergence. Numerical results support our findings.

***Index Terms*—** Large scale learning, Latent variable analysis, Expectation Maximization, Nonconvex stochastic optimization, Variance reduction.

1. INTRODUCTION

Intelligent processing of large data sets and efficient learning of high-dimensional models require new optimization algorithms designed to be robust to big data and complex models era (see e.g. [1–3]). This paper is concerned with stochastic optimization of a nonconvex finite-sum smooth objective function

$$\mathrm{Argmin}_{\theta\in\Theta} F(\theta), \qquad F(\theta) \stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) + \mathsf{R}(\theta) \;, \quad (1)$$

when $\Theta \subseteq \mathbb{R}^d$ and F cannot be explicitly evaluated (nor its gradient). Many statistical learning problems can be cast into this framework, where n is the number of observations or examples, $\mathcal{L}_i$ is a loss function associated to example $\#i$ (most often, a negative log-likelihood), and R is a penalty term promoting sparsity, regularity, etc. Empirical risk minimization in machine learning is a matter for (1). Intractability of $F(\theta)$ might come from two sources. The first, referred to as *large scale learning* setting, is that the number n is very large so that the computations involving a sum over n terms should be either simply avoided or sparingly used during the run of the optimization algorithm (see e.g. [4] for an introduction to the bridge between large scale learning and stochastic approximation; see [5,6] for applications to training of deep neural networks for signal and image processing). The second is due to the presence of latent variables: for any i, the function $\mathcal{L}_i$ is a (high-dimensional) integral over latent variables. Such a latent variable context is a classical statistical modeling: for example as a tool for solving inference in mixture models [7], for the definition of mixed models capturing variability among examples [8] or for modeling hidden and/or missing variables (see e.g. applications in text modeling through latent Dirichlet allocation [9], in audio source separation [10, 11], in hyper-spectral imaging [12]).

In this contribution, we address the two levels of intractability in the case $\mathcal{L}_i$ is of the form

$$\mathcal{L}_i(\theta) \stackrel{\mathrm{def}}{=} -\log \int_{\mathsf{Z}} h_i(z) \exp\left(\langle s_i(z), \phi(\theta)\rangle\right) \mu(\mathrm{d}z) \;. \quad (2)$$

This setting in particular covers the case when $\sum_{i=1}^n \mathcal{L}_i(\theta)$ is the negated log-likelihood of the observations $(Y_1, \cdots, Y_n)$, the pairs observation/latent variable $\{(Y_i, Z_i), i \le n\}$ are independent, and the distribution of the latent variable given the observation Y_i, given by $z \mapsto h_i(z) \exp\left(\langle s_i(z), \phi(\theta)\rangle\right) \mu(\mathrm{d}z)$ up to a multiplicative constant, is from the curved exponential family. Gaussian mixture models are typical examples, as well as mixtures of distributions from the curved exponential family. In the framework (1)-(2), a Majorize-Minimization approach through the Expectation-Maximization (EM) algorithm [13] is standard; unfortunately, the computational cost of the batch EM can be prohibitive in the large scale learning setting. Different strategies were proposed to address this issue [14–18]: they combine mini-batches processing, Stochastic Approximation (SA) techniques (see e.g. [19, 20]) and variance reduction methods.

The first contribution of this paper is to provide a novel algorithm, the generalized Stochastic Path-Integrated Differential EstimatoR EM (`g-SPIDER-EM`), which is among the variance reduced stochastic EM methods for nonconvex finite-sum optimization of the form (1)-(2); the generalizations allow a reduced computational cost without altering the convergence properties. The second contribution is the proof of complexity bounds, that is the number of parameter updates (M-step) and the number of conditional expectations evaluations (E-step), in order to reach ϵ-approximate stationary points; these bounds are derived for a specific form of `g-SPIDER-EM`: we show that the complexity bounds are the same as those of `SPIDER-EM`, bounds which are state of the art ones and overpass all the previous ones. Linear convergence rates are proved under a Polyak-Łojasiewicz condition. Finally, numerical results support our findings and provide insights on how to implement `g-SPIDER-EM` in order to inherit the properties of `SPIDER-EM` while reducing the computational cost.

Notations For $a, b \in \mathbb{R}^q$, $\langle a, b\rangle$ is the scalar product, and $\|\cdot\|$ the associated norm. For a matrix A, A^T is its transpose. For a positive integer n, set $[n]^\star \stackrel{\mathrm{def}}{=} \{1, \cdots, n\}$ and $[n] \stackrel{\mathrm{def}}{=} \{0, \cdots, n\}$. ∇f denotes the gradient of a differentiable function f. The minimum of a and b is denoted by $a \wedge b$. Finally, we use standard big O notation

Part of this work is funded by the Fondation Simone and Cino Del Duca under the program OpSiMorE

to leave out constants. For a random variable U and/or a filtration $\mathcal{F}$, $\sigma(U, \mathcal{F})$ denotes the sigma algebra generated by U and $\mathcal{F}$.

2. EM-BASED METHODS IN THE EXPECTATION SPACE

We begin by formulating the model assumptions:

A1. *$\Theta \subseteq \mathbb{R}^d$ is a convex set. $(\mathsf{Z}, \mathcal{Z})$ is a measurable space and μ is a σ-finite positive measure on $\mathcal{Z}$. The functions $\mathsf{R} : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$, $s_i : \mathsf{Z} \to \mathbb{R}^q$, $h_i : \mathsf{Z} \to \mathbb{R}_+$ for all $i \in [n]^\star$ are measurable. For any $\theta \in \Theta$ and $i \in [n]^\star$, $|\mathcal{L}_i(\theta)| < \infty$.*

For any $\theta \in \Theta$ and $i \in [n]^\star$, define the posterior density of the latent variable Z_i given the observation Y_i:

$$z \mapsto p_i(z;\theta) \stackrel{\text{def}}{=} h_i(z) \exp\left(\langle s_i(z), \phi(\theta)\rangle + \mathcal{L}_i(\theta)\right) , \quad (3)$$

note that the dependence upon y_i follows through the index i in the above. Set

$$\bar{s}_i(\theta) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} s_i(z)\, p_i(z;\theta)\mu(\mathrm{d}z), \quad \bar{s}(\theta) \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n \bar{s}_i(\theta) . \quad (4)$$

A2. *The expectations $\bar{s}_i(\theta)$ are well defined for all $\theta \in \Theta$ and $i \in [n]^\star$. For any $s \in \mathbb{R}^q$, $\mathrm{Argmin}_{\theta\in\Theta}\ (-\langle s, \phi(\theta)\rangle + \mathsf{R}(\theta))$ is a (non empty) singleton denoted by $\{\mathsf{T}(s)\}$.*

EM is an iterative algorithm: given a current value $\tau_k \in \Theta$, the next value is $\tau_{k+1} \leftarrow \mathsf{T} \circ \bar{s}(\tau_k)$. It combines an expectation step which boils down to the computation of $\bar{s}(\tau_k)$, an expectation under $p(\cdot;\tau_k)$; and a maximization step through the computation of the map T. Equivalently, by using T which maps $\mathbb{R}^q$ to Θ, it can be described in the *expectation space* (see [21]): given the current value $\bar{s}^k \in \bar{s}(\Theta)$, the next value is $\bar{s}^{k+1} \leftarrow \bar{s} \circ \mathsf{T}(\bar{s}^k)$. In this paper, we see EM as an iterative algorithm operating in the expectation space. In that case, the fixed points of the EM operator $\bar{s} \circ \mathsf{T}$ are the roots of the function h

$$h(s) \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(s) - s . \quad (5)$$

EM possesses a Lyapunov function: in the parameter space, it is the objective function F where by definition of the EM sequence, it holds $F(\tau_{k+1}) \le F(\tau_k)$; in the expectation space, it is $\mathrm{W} \stackrel{\text{def}}{=} F \circ \mathsf{T}$, and $\mathrm{W}(\bar{s}^{k+1}) \le \mathrm{W}(\bar{s}^k)$ holds. In order to derive complexity bounds, regularity assumptions are required on W:

A 3. *The functions ϕ and R are continuously differentiable on Θ^v, where Θ^v is an open neighborhood of Θ when Θ is not open and $\Theta^v \stackrel{\text{def}}{=} \Theta$ otherwise. T is continuously differentiable on $\mathbb{R}^q$. The function F is continuously differentiable on Θ^v and for any $\theta \in \Theta$, $\nabla F(\theta) = -\nabla\phi(\theta)^T \bar{s}(\theta) + \nabla\mathsf{R}(\theta)$. For any $s \in \mathbb{R}^q$, $B(s) \stackrel{\text{def}}{=} \nabla(\phi \circ \mathsf{T})(s)$ is a symmetric $q \times q$ matrix and there exist $0 < v_{min} \le v_{max} < \infty$ such that for all $s \in \mathbb{R}^q$, the spectrum of $B(s)$ is in $[v_{min}, v_{max}]$. For any $i \in [n]^\star$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i. The function $s \mapsto \nabla(F \circ \mathsf{T})(s) = -B(s)\,(\bar{s} \circ T(s) - s)$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_{W}.*

A3 implies that W has globally Lipschitz gradient and $\nabla W(s) = -B(s)h(s)$ for some positive definite matrix $B(s)$ (see e.g. [21, Lemma 2]; see also [22, Propositions 1 and 2]). Note that this implies that $\nabla W(s^\star) = 0$ iff $h(s^\star) = 0$.

Unfortunately, in the large scale learning setting (when $n \gg 1$), EM can not be easily applied since each iteration involves n conditional expectations evaluations through $\bar{s} = n^{-1}\sum_{i=1}^n \bar{s}_i$. Incremental EM techniques have been proposed to address this issue: the most straightforward approach amounts to use a SA scheme with mean field h since. Upon noting that $h(s) = \mathbb{E}\left[\bar{s}_I \circ \mathsf{T}(s)\right] - s$ where I is a uniform random variable (r.v.) on $[n]^\star$, the fixed points of the EM operator $\bar{s} \circ \mathsf{T}$ are those of the SA scheme

$$\widehat{S}_{k+1} = \widehat{S}_k + \gamma_{k+1}\left(\mathsf{b}^{-1}\sum_{i\in\mathcal{B}_{k+1}} \bar{s}_i \circ \mathsf{T}(\widehat{S}_k) - \widehat{S}_k\right) \quad (6)$$

where $\{\gamma_k, k \ge 0\}$ is a deterministic positive step size sequence, and $\mathcal{B}_{k+1}$ is sampled from $[n]^\star$ independently from the past of the algorithm. This forms the basis of `Online-EM` proposed by [15] (see also [23]). Variance reduced versions were also proposed and studied: Incremental EM (`i-EM`) [14, 24], Stochastic EM with variance reduction (`sEM-vr`) [16], Fast Incremental EM [17, 22] (`FIEM`) and more recently, Stochastic Path-Integrated Differential EstimatoR EM (`SPIDER-EM`) [18].

As shown in [22, section 2.3], these algorithms can be seen as a combination of SA with *control variate*: upon noting that $h(s) = h(s) + \mathbb{E}[U]$ for any r.v. U such that $\mathbb{E}[U] = 0$, *control variates within SA* procedures replace (6) with

$$\widehat{S}_{k+1} = \widehat{S}_k + \gamma_{k+1}\left(\mathsf{b}^{-1}\sum_{i\in\mathcal{B}_{k+1}} \bar{s}_i \circ \mathsf{T}(\widehat{S}_k) + U_{k+1} - \widehat{S}_k\right)$$

for a choice of U_{k+1} such that the new algorithm has better properties (for example, in terms of complexity - see the end of Section 3).

Lastly, we remark that A1–A3 are common assumptions (see e.g. [18] and [22] and references therein).

3. THE GEOM-SPIDER-EM ALGORITHM

Data: $k_{\text{out}} \in \mathbb{N}^\star$; $\widehat{S}_{\text{init}} \in \mathbb{R}^q$; $\xi_t \in \mathbb{N}^\star$ for $t \in [k_{\text{out}}]^\star$; $\gamma_{t,0} \ge 0$, $\gamma_{t,k} > 0$ for $t \in [k_{\text{out}}]^\star$, $k \in [\xi_t]^\star$.
Result: The g-SPIDER-EM sequence: $\{\widehat{S}_{t,k}\}$

$\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\text{init}}$;
$\mathsf{S}_{1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{1,-1}) + \mathcal{E}_1$;
for $t = 1, \cdots, k_{\text{out}}$ **do**
  for $k = 0, \ldots, \xi_t - 1$ **do**
    Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b from $[n]^\star$;
    $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \mathsf{b}^{-1}\sum_{i\in\mathcal{B}_{t,k+1}}\left(\bar{s}_i \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s}_i \circ \mathsf{T}(\widehat{S}_{t,k-1})\right)$;
    $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1}\left(\mathsf{S}_{t,k+1} - \widehat{S}_{t,k}\right)$
  $\widehat{S}_{t+1,-1} = \widehat{S}_{t,\xi_t}$;
  $\mathsf{S}_{t+1,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t+1,-1}) + \mathcal{E}_{t+1}$;
  $\widehat{S}_{t+1,0} = \widehat{S}_{t+1,-1} + \gamma_{t+1,0}\left(\mathsf{S}_{t+1,0} - \widehat{S}_{t+1,-1}\right)$

Algorithm 1: The `g-SPIDER-EM` algorithm. The $\mathcal{E}_t$'s are introduced as a perturbation to the computation of $\bar{s} \circ \mathsf{T}(\widehat{S}_{t,-1})$; they can be null.

The algorithm *generalized Stochastic Path-Integrated Differential EstimatoR Expectation Maximization* (`g-SPIDER-EM`) described by Algorithm 1 uses a new strategy when defining the approximation of $\bar{s} \circ \mathsf{T}(s)$ at each iteration. It is composed of nested loops: k_{out} outer loops, each of them formed with a possibly random number of inner loops. Within the tth outer loop, `g-SPIDER-EM` mimics the identity $\bar{s} \circ \mathsf{T}(\widehat{S}_{t,k}) = \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k-1}) + \{\bar{s} \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k-1})\}$. More precisely, at iteration $k+1$, the approximation

$\mathsf{S}_{t,k+1}$ of the full sum $\bar{s} \circ \mathsf{T}(\widehat{S}_{t,k})$ is the sum of the current approximation $\mathsf{S}_{t,k}$ and of a Monte Carlo approximation of the difference (see Lines 5, 6, in Algorithm 1); the examples i in $\mathcal{B}_{t,k+1}$ used in the approximation of $\bar{s} \circ \mathsf{T}(\widehat{S}_{t,k})$ and those used for the approximation of $\bar{s} \circ \mathsf{T}(\widehat{S}_{t,k-1})$ are the same - which make the approximations correlated and favor a variance reduction when plugged in the SA update (Line 7). $\mathcal{B}_{t,k+1}$ is sampled with or without replacement; even when $\mathcal{B}_{t,k+1}$ collects independent examples sampled uniformly from $[n]^\star$, we have $\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right] - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k}) = \mathsf{S}_{t,k} - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k-1})$ where $\mathcal{F}_{t,k}$ is the sigma-field collecting the randomness up to the end of the outer loop $\#t$ and inner loop $\#k$: the approximation $\mathsf{S}_{t,k+1}$ of $\bar{s} \circ \mathsf{T}(\widehat{S}_{t,k})$ is biased - a property which makes the theoretical analysis of the algorithm challenging. This approximation is reset (see Lines 2,9) at the end of an outer loop: in the "standard" `SPIDER-EM`, $\mathsf{S}_{t,0} = \bar{s} \circ \mathsf{T}(\widehat{S}_{t,-1})$ is computed, but this "refresh" can be only partial, by computing an update on a (large) batch $\tilde{\mathcal{B}}_{t,0}$ (size $\tilde{\mathsf{b}}_t$) of observations: $\mathsf{S}_{t,0} = \tilde{\mathsf{b}}_t^{-1} \sum_{i \in \tilde{\mathcal{B}}_{t,0}} \bar{s}_i \circ \mathsf{T}(\widehat{S}_{t,-1})$. Such a reset starts a so-called *epoch* (see Line 3). The number of inner loops ξ_t at epoch $\#t$ can be deterministic; or random, such as a uniform distribution on $[k_{\rm in}]^\star$ or a geometric distribution, and drawn prior the run of the algorithm.

Comparing `g-SPIDER-EM` with `SPIDER-EM` [18], we notice that *(i)* the former allows a perturbation $\mathcal{E}_t$ when initializing $\mathsf{S}_{t,0}$, which is important for computational cost reduction; *(ii)* `g-SPIDER-EM` considers epochs with time-varying length ξ_t which covers situations when it is random and chosen independently of the other sources of randomness (the errors $\mathcal{E}_t$, the mini batches $\mathcal{B}_{t,k+1}$). Hereafter, we provide an original analysis of an `g-SPIDER-EM`, namely `Geom-SPIDER-EM` which corresponds to the case $\xi_t \leftarrow \Xi_t, \Xi_t$ being a geometric r.v. on $\mathbb{N}^\star$ with success probability $1 - \rho_t \in (0,1)$: $\mathbb{P}(\Xi_t = k) = (1-\rho_t)\rho_t^{k-1}$ for $k \geq 1$ (hereafter, we will write $\Xi_t \sim \mathcal{G}^\star(1-\rho_t)$). Since Ξ_t is also the first success distribution in a sequence of independent Bernoulli trials, the geometric length could be replaced with: *(i)* at each iteration k of epoch t, sample a Bernoulli r.v. with a probability of success $(1-\rho_t)$; *(ii)* when the coin comes up head, start a new epoch (see [25,26] for similar ideas on stochastic gradient algorithms).

Let us establish complexity bounds for `Geom-SPIDER-EM`. We analyze a randomized terminating iteration $\Xi^\star$ [27] and discuss how to choose $k_{\rm out}, \mathsf{b}$ and $\xi_1, \cdots, \xi_{k_{\rm out}}$ as a function of the batch size n and an accuracy $\epsilon > 0$ to reach ϵ-approximate stationarity i.e. $\mathbb{E}[\|h(\widehat{S}_{\Xi^\star})\|^2] \leq \epsilon$. To this end, we endow the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with the sigma-fields $\mathcal{F}_{1,0} \overset{\rm def}{=} \sigma(\mathcal{E}_1)$, $\mathcal{F}_{t,0} \overset{\rm def}{=} \sigma(\mathcal{F}_{t-1,\xi_t}, \mathcal{E}_t)$ for $t \geq 2$, and $\mathcal{F}_{t,k+1} \overset{\rm def}{=} \sigma(\mathcal{F}_{t,k}, \mathcal{B}_{t,k+1})$ for $t \in [k_{\rm out}]^\star, k \in [\xi_t - 1]$. For a r.v. $\Xi_t \sim \mathcal{G}^\star(1-\rho_t)$, set $\overline{\mathbb{E}}_t[\phi(\Xi_t)|\mathcal{F}_{t,0}] \overset{\rm def}{=} (1-\rho_t)\sum_{k \geq 1} \rho_t^{k-1} \mathbb{E}[\phi(k)|\mathcal{F}_{t,0}]$ for any bounded measurable function ϕ.

Theorem 1. *Assume A1 to A3. For any $t \in [k_{\rm out}]^\star$, let $\rho_t \in (0,1)$ and $\Xi_t \sim \mathcal{G}^\star(1-\rho_t)$. Run Algorithm 1 with $\gamma_{t,k+1} = \gamma_t > 0$, $\gamma_{1,0} = 0$ and $\xi_t \leftarrow \Xi_t$ for any $t \in [k_{\rm out}]^\star$, $k \geq 0$. Then, for any $t \in [k_{\rm out}]^\star$,*

$$\begin{aligned}
&\frac{v_{\min}\gamma_t}{2(1-\rho_t)} \overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t-1})\|^2 | \mathcal{F}_{t,0}\right] \\
&\leq \mathrm{W}(\widehat{S}_{t,0}) - \overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right] + \frac{v_{\max}\gamma_t}{2(1-\rho_t)}\|\mathcal{E}_t\|^2 \\
&+ \frac{v_{\max}\gamma_t\gamma_{t,0}^2}{2(1-\rho_t)}\frac{L^2}{\mathsf{b}}\|\Delta\widehat{S}_{t,0}\|^2 + \mathcal{N}_t\,\overline{\mathbb{E}}_t\left[\|\Delta\widehat{S}_{t,\Xi}\|^2|\mathcal{F}_{t,0}\right] \;;
\end{aligned}$$

where $\Delta\widehat{S}_{t,\xi} \overset{\rm def}{=} \mathsf{S}_{t,\xi} - \widehat{S}_{t,\xi-1}$, $L^2 \overset{\rm def}{=} n^{-1}\sum_{i=1}^n L_i$, and

$$\mathcal{N}_t \overset{\rm def}{=} -\frac{\gamma_t}{2(1-\rho_t)}\left(v_{\min} - \gamma_t L_{\dot{\mathrm{W}}} - \frac{v_{\max}L^2\rho_t}{(1-\rho_t)\mathsf{b}}\gamma_t^2\right) .$$

Theorem 1 is the key result from which our conclusions are drawn; its proof is adapted from [18, section 8] (also see [28, Theorem 10]).

Let us discuss the rate of convergence and the complexity of `Geom-SPIDER-EM` in the case: for any $t \in [k_{\rm out}]^\star$, the mean number of inner loops is $(1-\rho_t)^{-1} = k_{\rm in}$, $\gamma_{t,0} = 0$ and $\gamma_t = \alpha/L$ for $\alpha > 0$ satisfying

$$v_{\min} - \alpha\frac{L_{\dot{\mathrm{W}}}}{L} - \alpha^2 v_{\max}\frac{k_{\rm in}}{\mathsf{b}}\left(1 - \frac{1}{k_{\rm in}}\right) > 0 .$$

Linear rate. When $\Xi \sim \mathcal{G}^\star(1-\rho)$, we have (see [28, Lemma 1])

$$\rho\mathbb{E}\left[D_\Xi\right] \leq \rho\mathbb{E}\left[D_\Xi\right] + (1-\rho)D_0 = \mathbb{E}\left[D_{\Xi-1}\right] \tag{7}$$

for any positive sequence $\{D_k, k \geq 0\}$; Theorem 1 implies

$$\begin{aligned}
\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t})\|^2|\mathcal{F}_{t,0}\right] &\leq \frac{2L}{v_{\min}\alpha(k_{\rm in}-1)}\left(\mathrm{W}(\widehat{S}_{t,0}) - \min \mathrm{W}\right) \\
&+ \frac{v_{\max}}{v_{\min}}\frac{k_{\rm in}}{k_{\rm in}-1}\|\mathcal{E}_t\|^2 \;; \quad (8)
\end{aligned}$$

see [28, Corollary 11]. Hence, when $\|\mathcal{E}_t\| = 0$ and W satisfies a Polyak-Łojasiewicz condition [29], i.e.

$$\exists \tau > 0, \forall s \in \mathbb{R}^q, \qquad \mathrm{W}(s) - \min \mathrm{W} \leq \tau\|\nabla \mathrm{W}(s)\|^2 \tag{9}$$

then (8) yields

$$\mathcal{H}_t \overset{\rm def}{=} \overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t})\|^2|\mathcal{F}_{t,0}\right] \leq \frac{2L\tau v_{\max}^2}{v_{\min}\alpha(k_{\rm in}-1)}\|h(\widehat{S}_{t-1,\Xi_{t-1}})\|^2 ,$$

thus establishing a linear rate of the algorithm along the path $\{\widehat{S}_{t,\Xi_t}, t \in [k_{\rm out}]^\star\}$ as soon as $k_{\rm in}$ is large enough:

$$\mathbb{E}\left[\mathcal{H}_t\right] \leq \left(\frac{2L\tau v_{\max}^2}{v_{\min}\alpha(k_{\rm in}-1)}\right)^t \|h(\widehat{S}_{\rm init})\|^2 .$$

Even if the Polyak-Łojasiewicz condition (9) is quite restrictive, the above discussion gives the intuition of the *lock-in* phenomenon which often happens at convergence: a linear rate of convergence is observed when the path is trapped in a neighborhood of its limiting point, which may be the consequence that locally, the Polyak-Łojasiewicz condition holds (see figure 1 in Section 4).

Complexity for ϵ-approximate stationarity. From Theorem 1, Eq. (7) and $\widehat{S}_{t,\Xi_t} = \widehat{S}_{t+1,0}$ (here $\gamma_{t,0} = 0$ and $\mathcal{E}_t = 0$), it holds

$$\frac{v_{\min}\alpha(k_{\rm in}-1)}{2L}\mathbb{E}\left[\mathcal{H}_t\right] \leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) - \mathrm{W}(\widehat{S}_{t+1,0})\right] .$$

Therefore,

$$\frac{1}{k_{\rm out}}\sum_{t=1}^{k_{\rm out}}\mathbb{E}\left[\mathcal{H}_t\right] \leq \frac{2L\left(\mathrm{W}(\widehat{S}_{\rm init}) - \min \mathrm{W}\right)}{v_{\min}\alpha(k_{\rm in}-1)k_{\rm out}} . \tag{10}$$

Eq. (10) establishes that in order to obtain an ϵ-approximate stationary point, it is sufficient to stop the algorithm at the end of the epoch $\#T$, where T is sampled uniformly from $[k_{\rm out}]^\star$ with $k_{\rm out} = O(L/(\epsilon\alpha k_{\rm in}))$ - and return $\widehat{S}_{T,\Xi_T}$. To do such, the mean number of conditional expectations evaluations is $\mathcal{K}_{\rm CE} \overset{\rm def}{=} n + nk_{\rm out} +$

$2\mathsf{b}k_{\rm in}k_{\rm out}$; and the mean number of optimization steps is $\mathcal{K}_{\rm Opt} \stackrel{\rm def}{=} k_{\rm out} + k_{\rm in}k_{\rm out}$. By choosing $k_{\rm in} = O(\sqrt{n})$ and $k_{\rm in}/\mathsf{b} = O(1)$, we have $\mathcal{K}_{\rm CE} = O(n + L\sqrt{n}/(\epsilon\alpha))$ and $\mathcal{K}_{\rm Opt} = O(L/(\epsilon\alpha))$. Similar randomized terminating strategies were proposed in the literature: their optimal complexity in terms of conditional expectations evaluations is $O(\epsilon^{-2})$ for `Online-EM` [15], $O(\epsilon^{-1}n)$ for `i-EM` [14], $O(\epsilon^{-1}n^{2/3})$ for `sEM-vr` [16,17], $O(\{\epsilon^{-1}n^{2/3}\} \wedge \{\epsilon^{-3/2}\sqrt{n}\})$ for `FIEM` [17,22] and $O(\epsilon^{-1}\sqrt{n})$ for `SPIDER-EM` - see [18, section 6] for a comparison of the complexities $\mathcal{K}_{\rm CE}$ and $\mathcal{K}_{\rm opt}$ of these incremental EM algorithms. Hence, `Geom-SPIDER-EM` has the same complexity bounds as `SPIDER-EM`, and they are optimal among the class of incremental EM algorithms.

4. NUMERICAL ILLUSTRATION

We perform experiments on the MNIST dataset, which consists of $n = 6 \times 10^4$ images of handwritten digits, each with 784 pixels. We pre-process the datas as detailed in [22, section 5]: 67 uninformative pixels are removed from each image, then a principal component analysis is applied to further reduce the dimension; we keep the 20 principal components of each observation. The learning problem consists in fitting a Gaussian mixture model with $g = 12$ components having the same covariance matrix: θ collects the weights of the mixture, the expectations of the components (i.e. g vectors in $\mathbb{R}^{20}$) and a 20×20 covariance matrix; F is the negative normalized log-likelihood (no penalty term). All the algorithms start from $\widehat{S}_{\rm init} = \bar{s} \circ \mathsf{T}(\theta_{\rm init})$ such that $-F(\theta_{\rm init}) = -58.3$, and their first two epochs are `Online-EM` iterations. The first epoch with a variance reduction technique is the 3rd; on Fig. 1, the plot starts at epoch $\#2$.

`Geom-SPIDER-EM` is run with a constant step size $\gamma_{t,k} = 0.01$ (and $\gamma_{t,0} = 0$); $k_{\rm out} = 148$ epochs (which are preceded with 2 epochs of `Online-EM`); a mini batch size $\mathsf{b} = \sqrt{n}$. Different strategies are considered for the initialization $\mathsf{S}_{t,0}$ and the parameter of the geometric r.v. Ξ_t. In `full-geom`, $k_{\rm in} = \sqrt{n}/2$ so that the mean total number of conditional expectations evaluations per outer loop is $2\mathsf{b}k_{\rm in} = n$; and $\mathcal{E}_t = 0$ which means that $\mathsf{S}_{t,0}$ requires the computation of the full sum $\bar{s}$ over n terms. In `half-geom`, $k_{\rm in}$ is defined as in `full-geom`, but for all $t \in [k_{\rm out}]^\star$, $\mathsf{S}_{t,0} = (2/n)\sum_{i\in\tilde{\mathcal{B}}_{t,0}} \bar{s}_i \circ \mathsf{T}(\widehat{S}_{t,-1})$ where $\mathcal{B}_{t,0}$ is of cardinality $n/2$; therefore $\mathcal{E}_t \neq 0$. In `quad-geom`, a quadratic growth is considered both for the mean of the geometric random variables: $\mathbb{E}[\Xi_t] = \min(n, \max(20t^2, n/50))/(2\mathsf{b})$; and for the size of the mini batch when computing $\mathsf{S}_{t,0}$: $\mathsf{S}_{t,0} = \tilde{\mathsf{b}}_t^{-1}\sum_{i\in\tilde{\mathcal{B}}_t} \bar{s}_i \circ \mathsf{T}(\widehat{S}_{t,-1})$ with $\tilde{\mathsf{b}}_t = \min(n, \max(20t^2, n/50))$. The `g-SPIDER-EM` with a constant number of inner loops $\xi_t = k_{\rm in} = n/(2\mathsf{b})$ is also run for comparison: different strategies for $\mathsf{S}_{t,0}$ are considered, the same as above (it corresponds to `full-ctt`, `half-ctt` and `quad-ctt` on the plots). Finally, in order to illustrate the benefit of the variance reduction, a pure `Online-EM` is run for 150 epochs, one epoch corresponding to $\sqrt{n}$ updates of the statistics $\widehat{S}$, each of them requiring a mini batch $\mathcal{B}_{k+1}$ of size $\sqrt{n}$ (see Eq.(6)).

The algorithms are compared through an estimation of the quantile of order 0.5 of $\|h(\widehat{S}_{t,\Xi_t})\|^2$ over 30 independent realizations. It is plotted versus the number of epochs t in Fig. 1 and the number of conditional expectations (CE) evaluations in Fig. 2. They are also compared through the objective function F along the path; the mean value over 30 independent paths is displayed versus the number of CE, see Fig. 3.

We first observe that `Online-EM` has a poor convergence rate, thus justifying the interest of variance reduction techniques as shown in Fig. 1. Having a persistent bias along iterations when defining $\mathsf{S}_{t,0}$

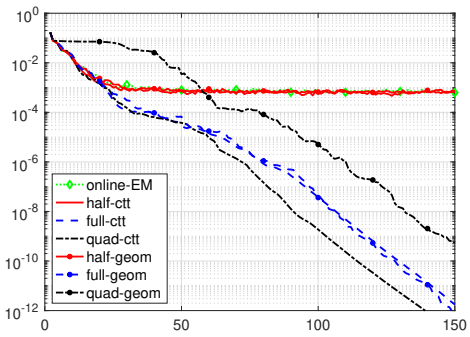


Fig. 1. Quantile 0.5 of $\|h(\widehat{S}_{t,\Xi_t})\|^2$ vs the number of epochs

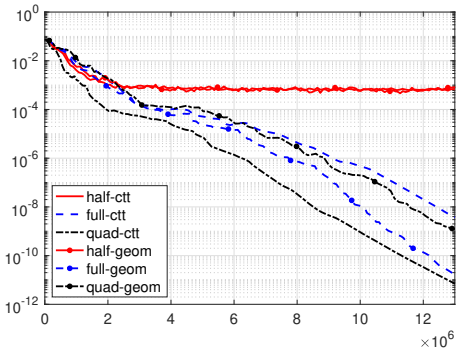


Fig. 2. Quantile 0.5 of $\|h(\widehat{S}_{t,\Xi_t})\|^2$ vs the number of CE evaluations

i.e. considering $\tilde{\mathsf{b}}_t \neq n$ and therefore $\mathcal{E}_t \neq 0$, is also a bad strategy as seen in Fig. 1, 2 for `half-ctt` and `half-geom`. For the four other `g-SPIDER-EM` strategies, we observe a linear convergence rate in Fig. 1, 2. The best strategy, both in terms of CE evaluations and in terms of efficiency given a number of epochs, is `quad-ctt`: a constant and deterministic number of inner loops ξ_t combined with an increasing accuracy when computing $\mathsf{S}_{t,0}$; therefore, during the first iterations, it is better to reduce the computational cost of the algorithm by considering $\tilde{\mathsf{b}}_t \ll n$. When $\mathcal{E}_t = 0$ (i.e. $\tilde{\mathsf{b}}_t = n$ so the computational cost of $\mathsf{S}_{t,0}$ is maximal), it is possible to reduce the total CE computational cost of the algorithm by considering a random number of inner loops (see `full-geom` and `full-ctt` on Fig. 1, 2). Finally, the strategy which consists in increasing both $\tilde{\mathsf{b}}_t$ and the number of inner loops, does not look the best one (see `quad-ctt` and `quad-geom` on Fig. 1 to Fig. 3).

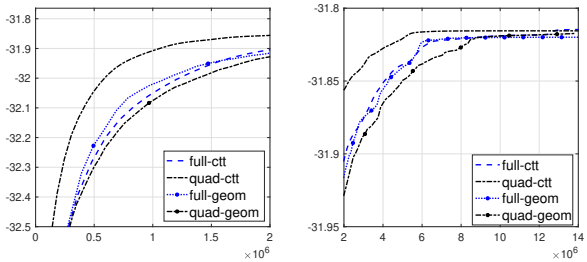


Fig. 3. (Left) $-F$ vs CE, until $2e6$. (Right) $-F$ vs CE, after $2e6$.

5. REFERENCES

[1] K. Slavakis, G.B. Giannakis, and G. Mateos, "Modeling and optimization for big data analytics: (statistical) learning tools for our era of data deluge," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 18–31, 2014.

[2] P. Bühlmann, P. Drineas, M. Kane, and M. van der Laan, *Handbook of Big Data*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2016.

[3] W. Härdle, H.H.S Lu, and X. Shen, *Handbook of big data analytics*, Springer, 2018.

[4] K. Slavakis, S. Kim, and G.B. Mateos, G. Giannakis, "Stochastic Approximation vis-a-vis Online Learning for Big Data Analytics [Lecture Notes]," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 124–129, 2014.

[5] Y. LeCun, B.H. Boser, J.S. Denker, D. Henderson, R.E. Howard, W.E. Hubbard, and L.D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzky, Ed., pp. 396–404. Morgan-Kaufmann, 1990.

[6] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] G.J. McLachlan and D. Peel, *Finite mixture models*, vol. 299 of *Probability and Statistics – Applied Probability and Statistics Section*, Wiley, New York, 2000.

[8] J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics. Springer, Dordrecht, 2007.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[10] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[11] K. Weisberg, S. Gannot, and O. Schwartz, "An online multiple-speaker doa tracking using the capp-moulines recursive expectation-maximization algorithm," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 656–660.

[12] B. Lin, X. Tao, S. Li, L. Dong, and J. Lu, "Variational bayesian image fusion based on combined sparse representations," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 1432–1436.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[14] R. M. Neal and G. E. Hinton, *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*, pp. 355–368, Springer Netherlands, Dordrecht, 1998.

[15] O. Cappé and E. Moulines, "On-line Expectation Maximization algorithm for latent data models," *J. Roy. Stat. Soc. B Met.*, vol. 71, no. 3, pp. 593–613, 2009.

[16] J. Chen, J. Zhu, Y.W. Teh, and T. Zhang, "Stochastic Expectation Maximization with Variance Reduction," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., pp. 7967–7977. 2018.

[17] B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle, "On the Global Convergence of (Fast) Incremental Expectation Maximization Methods," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds., pp. 2837–2847. Curran Associates, Inc., 2019.

[18] G. Fort, E. Moulines, and H.T. Wai, "A Stochastic Path-Integrated Differential EstimatoR Expectation Maximization Algorithm," in *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc., 2020.

[19] A. Benveniste, P. Priouret, and M. Métivier, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, Heidelberg, 1990.

[20] V. S. Borkar, *Stochastic approximation*, Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008, A dynamical systems viewpoint.

[21] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a Stochastic Approximation version of the EM algorithm," *Ann. Statist.*, vol. 27, no. 1, pp. 94–128, 1999.

[22] G. Fort, E. Moulines, and P. Gach, "Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence," Tech. Rep., HAL 02617725v2, 2020.

[23] P. Liang and D. Klein, "Online EM for unsupervised models," in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, pp. 611–619.

[24] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *J. Mach. Learn. Res.*, vol. 6, pp. 2049–2073, 2005.

[25] Z. Li, H. Bao, X. Zhang, and P. Richtrik, "PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization," Tech. Rep., arXiv 2008.10898, 2020.

[26] S. Horvath, L. Lei, P. Richtarik, and M.I. Jordan, "Adaptivity of Stochastic Gradient Methods for Nonconvex Optimization," Tech. Rep., arXiv 2002.05359, 2020.

[27] S. Ghadimi and G. Lan, "Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming," *SIAM J. Optimiz.*, vol. 23, no. 4, pp. 2341–2368, 2013.

[28] G. Fort, E. Moulines, and H.-T. Wai, "GEOM-SPIDER-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization," Tech. Rep., 2020, supplementary material, available at https://perso.math.univ-toulouse.fr/gfort/publications-2/conference-proceedings/ and at HAL-03021394.

[29] H. Karimi, J. Nutini, and M. Schmidt, "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 795–811.

SUPPLEMENTARY MATERIAL, PAPER "GEOM-SPIDER-EM: FASTER VARIANCE REDUCED STOCHASTIC EXPECTATION MAXIMIZATION FOR NONCONVEX FINITE-SUM OPTIMIZATION"

Gersende Fort[⋆] *Eric Moulines*[†] *Hoi-To Wai*[◇]

[⋆] Institut de Mathématiques de Toulouse, Université de Toulouse; CNRS UPS, F-31062 Toulouse Cedex, France
[†] Centre de Mathématiques Appliquées; Ecole Polytechnique; 91128 Palaiseau Cedex, France
[◇] Department of SEEM; The Chinese University of Hong Kong; Shatin, Hong Kong

The paper was accepted for publication in the Proceedings of the 46th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021).
This supplementary material was not part of the reviewing process; it is provided to make an explicit proof of the results claimed in the peer-reviewed paper.

1. PROOF OF THEOREM 1

Let $\{\mathcal{E}_t, t \in [k_{\rm out}]^\star\}$ and $\{\mathcal{B}_{t,k+1}, t \in [k_{\rm out}]^\star, k \in [\xi_t - 1]\}$ be random variables defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Define the filtrations $\mathcal{F}_{1,0} \stackrel{\rm def}{=} \sigma(\mathcal{E}_1)$, $\mathcal{F}_{t,0} \stackrel{\rm def}{=} \sigma(\mathcal{F}_{t-1,\xi_t}, \mathcal{E}_t)$ for $t \geq 2$, and $\mathcal{F}_{t,k+1} \stackrel{\rm def}{=} \sigma(\mathcal{F}_{t,k}, \mathcal{B}_{t,k+1})$ for $t \in [k_{\rm out}]^\star, k \in [\xi_t - 1]$.

For $\rho_t \in (0,1)$, set

$$\overline{\mathbb{E}}_t[\phi(\Xi_t)|\mathcal{F}_{t,0}] \stackrel{\rm def}{=} (1-\rho_t) \sum_{k \geq 1} \rho_t^{k-1} \mathbb{E}[\phi(k)|\mathcal{F}_{t,0}] \ ,$$

for any measurable positive function ϕ.

A1. *$\Theta \subseteq \mathbb{R}^d$ is a convex set. $(\mathsf{Z}, \mathcal{Z})$ is a measurable space and μ is a σ-finite positive measure on $\mathcal{Z}$. The functions $\mathsf{R} : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$, $s_i : \mathsf{Z} \to \mathbb{R}^q$, $h_i : \mathsf{Z} \to \mathbb{R}_+$ for all $i \in [n]^\star$ are measurable. For any $\theta \in \Theta$ and $i \in [n]^\star$, $|\mathcal{L}_i(\theta)| < \infty$.*

A2. *The expectations $\bar{s}_i(\theta)$ are well defined for all $\theta \in \Theta$ and $i \in [n]^\star$. For any $s \in \mathbb{R}^q$, $\mathrm{Argmin}_{\theta \in \Theta} \ (-\langle s, \phi(\theta) \rangle + \mathsf{R}(\theta))$ is a (non empty) singleton denoted by $\{\mathsf{T}(s)\}$.*

A 3. *The functions ϕ and R are continuously differentiable on Θ^v, where Θ^v is an open neighborhood of Θ when Θ is not open and $\Theta^v \stackrel{\rm def}{=} \Theta$ otherwise. T is continuously differentiable on $\mathbb{R}^q$. The function F is continuously differentiable on Θ^v and for any $\theta \in \Theta$, $\nabla F(\theta) = -\nabla\phi(\theta)^T \bar{s}(\theta) + \nabla \mathsf{R}(\theta)$. For any $s \in \mathbb{R}^q$, $B(s) \stackrel{\rm def}{=} \nabla(\phi \circ \mathsf{T})(s)$ is a symmetric $q \times q$ matrix and there exist $0 < v_{min} \leq v_{max} < \infty$ such that for all $s \in \mathbb{R}^q$, the spectrum of $B(s)$ is in $[v_{min}, v_{max}]$. For any $i \in [n]^\star$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i. The function $s \mapsto \nabla(F \circ \mathsf{T})(s) = -B(s)\,(\bar{s} \circ T(s) - s)$ is globally Lipschitz on $\mathbb{R}^q$ with constant $L_{\dot{\rm W}}$.*

Lemma 1. *Let $\rho \in (0,1)$ and $\{D_k, k \geq 0\}$ be real numbers such that $\sum_{k \geq 0} \rho^k |D_k| < \infty$. Let $\xi \sim \mathcal{G}^\star(1-\rho)$. Then $\mathbb{E}[D_{\xi-1}] = \rho\mathbb{E}[D_\xi] + (1-\rho)D_0 = \mathbb{E}[D_\xi] + (1-\rho)(D_0 - \mathbb{E}[D_\xi])$.*

Part of this work is funded by the Fondation Simone and Cino Del Duca under the program OpSiMorE

Proof. By definition of ξ,

$$\begin{aligned}
\mathbb{E}\left[D_\xi\right] &= (1-\rho) \sum_{k \geq 1} \rho^{k-1} D_k \\
&= \rho^{-1}(1-\rho) \sum_{k \geq 1} \rho^k D_k \\
&= \rho^{-1}(1-\rho) \sum_{k \geq 0} \rho^k D_k - \rho^{-1}(1-\rho) D_0 \\
&= \rho^{-1}(1-\rho) \sum_{k \geq 1} \rho^{k-1} D_{k-1} - \rho^{-1}(1-\rho) D_0 \\
&= \rho^{-1}\mathbb{E}\left[D_{\xi-1}\right] - \rho^{-1}(1-\rho) D_0 \ .
\end{aligned}$$

This yields $\rho\mathbb{E}\left[D_\xi\right] = \mathbb{E}\left[D_{\xi-1}\right] - (1-\rho)D_0$ and concludes the proof. □

Lemma 2. *For any $t \in [k_{\rm out}]^\star$, $k \in [\xi_t]^\star$, $\mathcal{B}_{t,k}$ and $\mathcal{F}_{t,k-1}$ are independent. In addition, for any $s \in \mathbb{R}^q$, $\mathsf{b}^{-1}\mathbb{E}\left[\sum_{i \in \mathcal{B}_{t,k}} \bar{s}_i \circ \mathsf{T}(s)\right] = \bar{s} \circ \mathsf{T}(s)$. Finally, assume that for any $i \in [n]^\star$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i. Then for any $s, s' \in \mathbb{R}^q$,*

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k}} \left\{\bar{s}_i \circ \mathsf{T}(s) - \bar{s}_i \circ \mathsf{T}(s')\right\} - \bar{s} \circ \mathsf{T}(s) + \bar{s} \circ \mathsf{T}(s')\right\|^2\right] \\
\leq \frac{1}{\mathsf{b}} \left(L^2 \|s - s'\|^2 - \|\bar{s} \circ \mathsf{T}(s) - \bar{s} \circ \mathsf{T}(s')\|^2\right) \ ,
\end{aligned}$$

where $L^2 \stackrel{\rm def}{=} n^{-1} \sum_{i=1}^n L_i^2$.

Proof. See [1, Lemma 4]; the proof holds true when $\mathcal{B}_{t,k}$ is sampled with or without replacement. □

Proposition 3. *For any $t \in [k_{\rm out}]^\star$, $k \in [\xi_t - 1]$,*

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} | \mathcal{F}_{t,k}\right] - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k}) = \mathsf{S}_{t,k} - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k-1}) \ ,$$

and

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k}) | \mathcal{F}_{t,0}\right] = \mathcal{E}_t \ .$$

Proof. Let $t \in [k_{\rm out}]^\star$, $k \in [\xi_t - 1]$. By Lemma 2,

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} | \mathcal{F}_{t,k}\right] = \mathsf{S}_{t,k} + \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,k-1}) \ .$$

By definition of $\mathsf{S}_{t,0}$ and of the filtrations, $\mathsf{S}_{t,0} - \bar{s} \circ \mathsf{T}(\widehat{S}_{t,-1}) = \mathcal{E}_t \in \mathcal{F}_{t,0}$. The proof follows by induction on k. □

Proposition 4. *Assume that for any $i \in [n]^\star$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i. For any $t \in [k_{\rm out}]^\star$, $k \in [\xi_t - 1]$,*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right]\|^2|\mathcal{F}_{t,k}\right] \\
&\le \frac{1}{\mathsf{b}}\left(L^2\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2 - \|\bar{s}\circ\mathsf{T}(\widehat{S}_{t,k}) - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1})\|^2\right) \\
&\le \frac{L^2}{\mathsf{b}}\gamma_{t,k}^2\|\mathsf{S}_{t,k} - \widehat{S}_{t,k-1}\|^2 \ ,
\end{aligned}$$

where $L^2 \stackrel{\rm def}{=} n^{-1}\sum_{i=1}^n L_i^2$. By convention, $\gamma_{1,0} = 0$.

Proof. Let $t \in [k_{\rm out}]^\star$, $k \in [\xi_t - 1]$. By Lemma 2, Proposition 3, the definition of $\mathsf{S}_{t,k+1}$ and of the filtration $\mathcal{F}_{t,k}$,

$$\begin{aligned}
&\mathsf{S}_{t,k+1} - \mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right] \\
&= \mathsf{S}_{t,k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k}) - \mathsf{S}_{t,k} + \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1}) \\
&= \mathsf{b}^{-1}\sum_{i\in\mathcal{B}_{t,k+1}}\{\bar{s}_i\circ\mathsf{T}(\widehat{S}_{t,k}) - \bar{s}_i\circ\mathsf{T}(\widehat{S}_{t,k-1})\} \\
&\quad - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k}) + \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1}) \ .
\end{aligned}$$

We then conclude by Lemma 2 for the first inequality; and by using the definition of $\widehat{S}_{t,k}$ for the second one:

$$\widehat{S}_{t,k} - \widehat{S}_{t,k-1} = \gamma_{t,k}\left(\mathsf{S}_{t,k} - \widehat{S}_{t,k-1}\right)$$

except when $t = 1$ and $k = 0$, where $\widehat{S}_{1,0} - \widehat{S}_{1,-1} = 0$. □

Proposition 5. *Assume that for any $i \in [n]^\star$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i. For any $t \in [k_{\rm out}]^\star$, $k \in [\xi_t - 1]$,*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,k}\right] \\
&\le \frac{L^2}{\mathsf{b}}\gamma_{t,k}^2\|\mathsf{S}_{t,k} - \widehat{S}_{t,k-1}\|^2 + \|\mathsf{S}_{t,k} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1})\|^2 \ ,
\end{aligned}$$

where $L^2 \stackrel{\rm def}{=} n^{-1}\sum_{i=1}^n L_i^2$. By convention, $\gamma_{1,0} = 0$.

Proof. By definition of the conditional expectation, we have for any r.v. $\phi(V)$

$$\mathbb{E}\left[\|U - \phi(V)\|^2|V\right] = \mathbb{E}\left[\|U - \mathbb{E}\left[U|V\right]\|^2|V\right] + \|\mathbb{E}\left[U|V\right] - \phi(V)\|^2 \ .$$

The proof follows from this equality and Propositions 3 and 4. □

Corollary 6. *Assume that for any $i \in [n]^\star$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz with constant L_i. For any $t \in [k_{\rm out}]^\star$, let $\rho_t \in (0,1)$ and $\Xi_t \sim \mathcal{G}^\star(1-\rho_t)$. For any $t \in [k_{\rm out}]^\star$,*

$$\begin{aligned}
&\overline{\mathbb{E}}_t\left[\left(\gamma_{t,\Xi_t} - \rho_t\gamma_{t,\Xi_t+1}\right)\|\mathsf{S}_{t,\Xi_t} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right] \\
&\le \frac{L^2\rho_t}{\mathsf{b}}\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t+1}\gamma_{t,\Xi_t}^2\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right] \\
&\quad + \frac{L^2(1-\rho_t)}{\mathsf{b}}\gamma_{t,1}\gamma_{t,0}^2\|\mathsf{S}_{t,0} - \widehat{S}_{t,-1}\|^2 + (1-\rho_t)\gamma_{t,1}\|\mathcal{E}_t\|^2 \ ,
\end{aligned}$$

where $L^2 \stackrel{\rm def}{=} n^{-1}\sum_{i=1}^n L_i^2$.

Proof. Let $t \in [k_{\rm out}]^\star$ and $k \in [\xi_t - 1]$. From Proposition 5 and since $\mathcal{F}_{t,0} \subseteq \mathcal{F}_{t,k}$ for $k \in [\xi_t - 1]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,0}\right] \\
&\le \mathbb{E}\left[\|\mathsf{S}_{t,k} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k-1})\|^2 + \frac{L^2}{\mathsf{b}}\gamma_{t,k}^2\|\mathsf{S}_{t,k} - \widehat{S}_{t,k-1}\|^2|\mathcal{F}_{t,0}\right] \ .
\end{aligned}$$

Multiply by $\gamma_{t,k+1}$ and apply with $k = \xi_t - 1$:

$$\begin{aligned}
&\mathbb{E}\left[\gamma_{t,\xi_t}\|\mathsf{S}_{t,\xi_t} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\xi_t-1})\|^2|\mathcal{F}_{t,0}\right] \\
&\le \mathbb{E}\left[\gamma_{t,\xi_t}\|\mathsf{S}_{t,\xi_t-1} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\xi_t-2})\|^2|\mathcal{F}_{t,0}\right] \\
&\quad + \frac{L^2}{\mathsf{b}}\mathbb{E}\left[\gamma_{t,\xi_t}\gamma_{t,\xi_t-1}^2\|\mathsf{S}_{t,\xi_t-1} - \widehat{S}_{t,\xi_t-2}\|^2|\mathcal{F}_{t,0}\right] \ .
\end{aligned}$$

This implies

$$\begin{aligned}
&\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t}\|\mathsf{S}_{t,\Xi_t} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right] \\
&\le \overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t}\|\mathsf{S}_{t,\Xi_t-1} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-2})\|^2|\mathcal{F}_{t,0}\right] \\
&\quad + \frac{L^2}{\mathsf{b}}\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t}\gamma_{t,\Xi_t-1}^2\|\mathsf{S}_{t,\Xi_t-1} - \widehat{S}_{t,\Xi_t-2}\|^2|\mathcal{F}_{t,0}\right] \ .
\end{aligned}$$

By Lemma 1, we have

$$\begin{aligned}
&\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t}\|\mathsf{S}_{t,\Xi_t-1} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-2})\|^2|\mathcal{F}_{t,0}\right] \\
&= \rho_t\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t+1}\|\mathsf{S}_{t,\Xi_t} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right] \\
&\quad + (1-\rho_t)\gamma_{t,1}\|\mathsf{S}_{t,0} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,-1})\|^2 \ ;
\end{aligned}$$

by definition of $\mathsf{S}_{t,0}$ and $\mathcal{F}_{t,0}$, the last term is equal to $(1-\rho_t)\gamma_{t,1}\|\mathcal{E}_t\|^2$. By Lemma 1, we have

$$\begin{aligned}
&\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t}\gamma_{t,\Xi_t-1}^2\|\mathsf{S}_{t,\Xi_t-1} - \widehat{S}_{t,\Xi_t-2}\|^2|\mathcal{F}_{t,0}\right] \\
&= \rho_t\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t+1}\gamma_{t,\Xi_t}^2\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right] \\
&\quad + (1-\rho_t)\gamma_{t,1}\gamma_{t,0}^2\|\mathsf{S}_{t,0} - \widehat{S}_{t,-1}\|^2 \ .
\end{aligned}$$

This concludes the proof. □

Lemma 7. *For any $h, s, S \in \mathbb{R}^q$ and any $q \times q$ symmetric matrix B, it holds*

$$-2\left\langle Bh, S\right\rangle = -\left\langle BS, S\right\rangle - \left\langle Bh, h\right\rangle + \left\langle B\{h - S\}, h - S\right\rangle \ .$$

Proposition 8. *Assume A1 to A3. For any $t \in [k_{\rm out}]^\star$ and $k \in [\xi_t - 1]$,*

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k+1})|\mathcal{F}_{t,0}\right] + \frac{v_{\min}}{2}\gamma_{t,k+1}\mathbb{E}\left[\|h(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,0}\right] \\
&\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] \\
&\quad + \frac{v_{\max}}{2}\gamma_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,0}\right] \\
&\quad - \frac{\gamma_{t,k+1}}{2}\left(v_{\min} - \gamma_{t,k+1}L_{\dot{\mathrm{W}}}\right)\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \widehat{S}_{t,k}\|^2|\mathcal{F}_{t,0}\right] \ .
\end{aligned}$$

Proof. Since W is continuously differentiable with $L_{\dot{\mathrm{W}}}$-Lipschitz gradient, then for any $s, s' \in \mathbb{R}^q$,

$$\mathrm{W}(s') - \mathrm{W}(s) \le \left\langle\nabla\,\mathrm{W}(s), s' - s\right\rangle + \frac{L_{\dot{\mathrm{W}}}}{2}\|s' - s\|^2 \ .$$

Set $s' = s + \gamma S$ where $\gamma > 0$ and $S \in \mathbb{R}^q$. Since $\nabla\,\mathrm{W}(s) = -B(s)h(s)$ and $B(s)$ is symmetric, apply Lemma 7 with $h \leftarrow h(s)$, $B \leftarrow B(s)$ and $S = (s' - s)/\gamma$; this yields

$$\begin{aligned}
&\mathrm{W}(s + \gamma S) - \mathrm{W}(s) \le -\frac{\gamma}{2}\left\langle B(s)S, S\right\rangle - \frac{\gamma}{2}\left\langle B(s)h(s), h(s)\right\rangle \\
&\quad + \frac{\gamma}{2}\left\langle B(s)\{h(s) - S\}, h(s) - S\right\rangle + \frac{L_{\dot{\mathrm{W}}}}{2}\gamma^2\|S\|^2 \ .
\end{aligned}$$

Since $\|a\|^2 v_{\min} \leq \langle B(s)a, a\rangle \leq v_{\max}\|a\|^2$ for any $a \in \mathbb{R}^q$, we have

$$\mathrm{W}(s+\gamma S) - \mathrm{W}(s) \leq -\frac{\gamma v_{\min}}{2}\|S\|^2 - \frac{\gamma v_{\min}}{2}\|h(s)\|^2 + \frac{\gamma v_{\max}}{2}\|h(s)-S\|^2 + \frac{L_{\dot{\mathrm{W}}}}{2}\gamma^2\|S\|^2 \ .$$

Let $t \in [k_{\mathrm{out}}]^\star$ and $k \in [\xi_t - 1]$. Applying this inequality with $s \leftarrow \widehat{S}_{t,k}$, $\gamma \leftarrow \gamma_{t,k+1}$, $S \leftarrow \mathsf{S}_{t,k+1} - \widehat{S}_{t,k}$ (which yields $s + \gamma S = \widehat{S}_{t,k+1}$), and then the conditional expectation yield the result. □

Proposition 9. *Assume A1 to A3. For any $t \in [k_{\mathrm{out}}]^\star$*

$$\begin{aligned}&\mathrm{W}(\widehat{S}_{t+1,0}) - \mathrm{W}(\widehat{S}_{t+1,-1})\\ &\leq -\frac{\gamma_{t+1,0} v_{\min}}{2}\|h(\widehat{S}_{t+1,-1})\|^2 + \frac{v_{\max}\gamma_{t+1,0}}{2}\|\mathcal{E}_{t+1}\|^2\\ &- \frac{\gamma_{t+1,0}}{2}\left(v_{\min} - \gamma_{t+1,0}L_{\dot{\mathrm{W}}}\right)\|\mathsf{S}_{t+1,0} - \widehat{S}_{t+1,-1}\|^2 \ .\end{aligned}$$

Proof. As in the proof of Proposition 8, we write for any $s, s' \in \mathbb{R}^q$,

$$\mathrm{W}(s') - \mathrm{W}(s) \leq \left\langle \nabla \mathrm{W}(s), s' - s\right\rangle + \frac{L_{\dot{\mathrm{W}}}}{2}\|s'-s\|^2 \ .$$

With Lemma 7, this yields when $s' = s + \gamma S$ for $\gamma > 0$ and $S \in \mathbb{R}^q$

$$\mathrm{W}(s+\gamma S) - \mathrm{W}(s) \leq -\frac{\gamma}{2}\left(v_{\min} - \gamma L_{\dot{\mathrm{W}}}\right)\|S\|^2 - \frac{\gamma v_{\min}}{2}\|h(s)\|^2 + \frac{v_{\max}\gamma}{2}\|h(s)-S\|^2 \ .$$

Apply this inequality with $\gamma \leftarrow \gamma_{t+1,0}$, $s \leftarrow \widehat{S}_{t+1,-1}$ and $S \leftarrow \mathsf{S}_{t+1,0} - \widehat{S}_{t+1,-1}$. This yields $s + \gamma S = \widehat{S}_{t+1,0}$ and

$$h(s) - S = \bar{s} \circ \mathsf{T}(\widehat{S}_{t+1,-1}) - \mathsf{S}_{t+1,0} = -\mathcal{E}_{t+1} \ .$$

□

Theorem 10. *Assume A1 to A3. For any $t \in [k_{\mathrm{out}}]^\star$, let $\rho_t \in (0,1)$ and $\Xi_t \sim \mathcal{G}^\star(1-\rho_t)$. Finally, choose $\gamma_{t,k+1} = \gamma_t > 0$ for any $k \geq 0$. For any $t \in [k_{\mathrm{out}}]^\star$,*

$$\begin{aligned}&\frac{v_{\min}\gamma_t}{2(1-\rho_t)}\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right] \leq \mathrm{W}(\widehat{S}_{t,0}) - \overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right]\\ &+ \frac{v_{\max}}{2(1-\rho_t)}\frac{L^2}{\mathsf{b}}\gamma_t\gamma_{t,0}^2\|\mathsf{S}_{t,0} - \widehat{S}_{t,-1}\|^2 + \frac{v_{\max}}{2(1-\rho_t)}\gamma_t\|\mathcal{E}_t\|^2\\ &- \frac{\gamma_t}{2(1-\rho_t)}\left(v_{\min} - \gamma_t L_{\dot{\mathrm{W}}} - \frac{v_{\max}L^2\rho_t}{(1-\rho_t)\mathsf{b}}\gamma_t^2\right)\cdots\\ &\quad\times\overline{\mathbb{E}}_t\left[\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right] \ .\end{aligned}$$

By convention, $\gamma_{1,0} = 0$.

Proof. Apply Proposition 8 with $k \leftarrow \xi_t - 1$ and then set $\xi_t \leftarrow \Xi_t$; this yields

$$\begin{aligned}&\overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right] + \frac{v_{\min}}{2}\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t}\|h(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right]\\ &\leq \overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t-1})|\mathcal{F}_{t,0}\right]\\ &+ \frac{v_{\max}}{2}\overline{\mathbb{E}}_t\left[\gamma_{t,\Xi_t}\|\mathsf{S}_{t,\Xi_t} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right]\\ &- \overline{\mathbb{E}}_t\left[\frac{\gamma_{t,\Xi_t}}{2}\left(v_{\min} - \gamma_{t,\Xi_t}L_{\dot{\mathrm{W}}}\right)\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right] \ .\end{aligned}$$

Since $\Xi_t \geq 1$ and $\gamma_{t,k} = \gamma_t$ for any $k \geq 1$, we have

$$\begin{aligned}&\overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right] + \frac{v_{\min}}{2}\gamma_t\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right]\\ &\leq \overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t-1})|\mathcal{F}_{t,0}\right]\\ &+ \frac{v_{\max}}{2}\gamma_t\overline{\mathbb{E}}_t\left[\|\mathsf{S}_{t,\Xi_t} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right]\\ &- \frac{\gamma_t}{2}\left(v_{\min} - \gamma_t L_{\dot{\mathrm{W}}}\right)\overline{\mathbb{E}}_t\left[\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right] \ .\end{aligned}$$

By Lemma 1, it holds

$$\begin{aligned}&\overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right] = \overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t-1})|\mathcal{F}_{t,0}\right]\\ &+ (1-\rho_t)\left(\overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right] - \mathrm{W}(\widehat{S}_{t,0})\right) \ .\end{aligned}$$

Furthermore, by Corollary 6 applied with $\gamma_{t,\Xi_t} = \gamma_{t,\Xi_t+1} = \gamma_t$

$$\begin{aligned}&(1-\rho_t)\gamma_t\overline{\mathbb{E}}_t\left[\|\mathsf{S}_{t,\Xi_t} - \bar{s}\circ\mathsf{T}(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right]\\ &\leq \frac{L^2\rho_t}{\mathsf{b}}\gamma_t^3\overline{\mathbb{E}}_t\left[\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right]\\ &+ \frac{L^2(1-\rho_t)}{\mathsf{b}}\gamma_t\gamma_{t,0}^2\|\mathsf{S}_{t,0} - \widehat{S}_{t,-1}\|^2 + (1-\rho_t)\gamma_t\|\mathcal{E}_t\|^2 \ ,\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{v_{\min}}{2}\gamma_t\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right]\\ &\leq (1-\rho_t)\,\mathrm{W}(\widehat{S}_{t,0}) - (1-\rho_t)\overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right]\\ &+ \frac{v_{\max}}{2}\frac{L^2\rho_t}{(1-\rho_t)\mathsf{b}}\gamma_t^3\overline{\mathbb{E}}_t\left[\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right]\\ &+ \frac{v_{\max}}{2}\frac{L^2}{\mathsf{b}}\gamma_t\gamma_{t,0}^2\|\mathsf{S}_{t,0} - \widehat{S}_{t,-1}\|^2 + \frac{v_{\max}}{2}\gamma_t\|\mathcal{E}_t\|^2\\ &- \frac{\gamma_t}{2}\left(v_{\min} - \gamma_t L_{\dot{\mathrm{W}}}\right)\overline{\mathbb{E}}_t\left[\|\mathsf{S}_{t,\Xi_t} - \widehat{S}_{t,\Xi_t-1}\|^2|\mathcal{F}_{t,0}\right] \ .\end{aligned}$$

This concludes the proof. □

Corollary 11 (of Theorem 10)**.** *Assume that for any $t \in [k_{\mathrm{out}}]^\star$, $1-\rho_t = 1/k_{\mathrm{in}}$ and $\gamma_t = \alpha/L$ where $\alpha > 0$ satisfies*

$$v_{\min} - \alpha\frac{L_{\dot{\mathrm{W}}}}{L} - \alpha^2 v_{\max}\frac{k_{\mathrm{in}}}{\mathsf{b}}\left(1 - \frac{1}{k_{\mathrm{in}}}\right) > 0 \ .$$

For any $t \in [k_{\mathrm{out}}]^\star$,

$$\begin{aligned}&\left(\frac{\alpha(k_{\mathrm{in}}-1)}{L} + \gamma_{t+1,0}\right)\frac{v_{\min}}{2}\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t})\|^2|\mathcal{F}_{t,0}\right]\\ &\leq \mathrm{W}(\widehat{S}_{t,0}) - \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t+1,0})|\mathcal{F}_{t,0}\right]\\ &- \frac{\gamma_{t+1,0}}{2}\left(v_{\min} - \gamma_{t+1,0}L_{\dot{\mathrm{W}}}\right)\mathbb{E}\left[\|\mathsf{S}_{t+1,0} - \widehat{S}_{t+1,-1}\|^2|\mathcal{F}_{t,0}\right]\\ &+ \frac{v_{\max}k_{\mathrm{in}}}{2}\frac{\alpha L}{\mathsf{b}}\gamma_{t,0}^2\|\mathsf{S}_{t,0} - \widehat{S}_{t,-1}\|^2\\ &+ \frac{v_{\max}\alpha k_{\mathrm{in}}}{2L}\|\mathcal{E}_t\|^2 + \frac{v_{\max}\gamma_{t+1,0}}{2}\mathbb{E}\left[\|\mathcal{E}_{t+1}\|^2|\mathcal{F}_{t,0}\right] \ .\end{aligned}$$

Proof. Let $t \in [k_{\rm out}]^\star$. By Proposition 9, since $\widehat{S}_{t,\xi_t} = \widehat{S}_{t+1,-1}$ we have

$$
\begin{aligned}
&- \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,\xi_t})|\mathcal{F}_{t,0}\right] \leq -\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t+1,0})|\mathcal{F}_{t,0}\right] \\
&- \frac{\gamma_{t+1,0}\upsilon_{\min}}{2}\mathbb{E}\left[\|h(\widehat{S}_{t,\xi_t})\|^2|\mathcal{F}_{t,0}\right] \\
&+ \frac{\upsilon_{\max}\gamma_{t+1,0}}{2}\mathbb{E}\left[\|\mathcal{E}_{t+1}\|^2|\mathcal{F}_{t,0}\right] \\
&- \frac{\gamma_{t+1,0}}{2}\left(\upsilon_{\min} - \gamma_{t+1,0}L_{\dot{\mathrm{W}}}\right)\mathbb{E}\left[\|\mathsf{S}_{t+1,0} - \widehat{S}_{t+1,-1}\|^2|\mathcal{F}_{t,0}\right] \ .
\end{aligned}
$$

The previous inequality remains true when $\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,\xi_t})|\mathcal{F}_{t,0}\right]$ is replaced with $\overline{\mathbb{E}}_t\left[\mathrm{W}(\widehat{S}_{t,\Xi_t})|\mathcal{F}_{t,0}\right]$; and $\mathbb{E}\left[\|h(\widehat{S}_{t,\xi_t})\|^2|\mathcal{F}_{t,0}\right]$ with $\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t})\|^2|\mathcal{F}_{t,0}\right]$. The proof follows from Theorem 10, and (see Lemma 1)

$$
\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t-1})\|^2|\mathcal{F}_{t,0}\right] \geq \rho_t\overline{\mathbb{E}}_t\left[\|h(\widehat{S}_{t,\Xi_t})\|^2|\mathcal{F}_{t,0}\right] \ .
$$

We also use $\gamma_t = \alpha/L$ and $\rho_t/(1-\rho_t) = k_{\rm in} - 1$. □

2. NUMERICAL ILLUSTRATION

By convention, vectors are column-vectors. For a matrix a, a^T denotes its transpose.

2.1. The model

The observations $(Y_1, \cdots, Y_n)$ are assumed i.i.d. with distribution

$$\sum_{\ell=1}^{g} \alpha_\ell \, \mathcal{N}_p(\mu_\ell, \Sigma)$$

where $\alpha_\ell \geq 0$ and $\sum_{\ell=1}^g \alpha_\ell = 1$. The negative normalized log-likelihood is given by (up to an additive constant)

$$\theta \mapsto F(\theta) \stackrel{\text{def}}{=} -\frac{1}{2}\log\det(\Sigma^{-1}) - \frac{1}{n}\sum_{i=1}^{n} \log \sum_{\ell=1}^{g} \alpha_\ell \exp\left(-\frac{1}{2}(Y_i-\mu_\ell)^T\Sigma^{-1}(Y_i-\mu_\ell)\right).$$

The parameter θ collects the weights, the expectations, and the covariance matrix

$$\theta \stackrel{\text{def}}{=} (\alpha_1, \cdots, \alpha_g, \mu_1, \cdots, \mu_g, \Sigma)$$

so that $\theta \in \Theta$ with

$$\Theta \stackrel{\text{def}}{=} \left\{ \alpha_\ell \geq 0 \text{ for } \ell \in [g]^\star \text{ s.t. } \sum_{\ell=1}^{g} \alpha_\ell = 1 \right\} \times (\mathbb{R}^p)^g \times \mathcal{M}_p^+;$$

$\mathcal{M}_p^+$ denotes the set of the $p \times p$ positive definite matrices.

For all $i \in [n]^\star$, we write

$$\begin{aligned}
&-\log \sum_{\ell=1}^{g} \alpha_\ell \exp\left(-\frac{1}{2}(Y_i-\mu_\ell)^T\Sigma^{-1}(Y_i-\mu_\ell)\right) \\
&= -\log \sum_{\ell=1}^{g} \exp\left(-\frac{1}{2}(Y_i-\mu_\ell)^T\Sigma^{-1}(Y_i-\mu_\ell) + \log\alpha_\ell\right) \\
&= -\log \sum_{z=1}^{g} \exp\left(-\frac{1}{2}\sum_{\ell=1}^{g} \mathsf{1}_{z=\ell}(Y_i-\mu_\ell)^T\Sigma^{-1}(Y_i-\mu_\ell) \right. \\
&\left. + \sum_{\ell=1}^{g} \mathsf{1}_{z=\ell}\log\alpha_\ell\right) \\
&= \frac{1}{2}Y_i^T\Sigma^{-1}Y_i - \log\sum_{z=1}^{g}\exp\left(\sum_{\ell=1}^{g}\mathsf{1}_{z=\ell}Y_i^T\Sigma^{-1}\mu_\ell\right)\cdots \\
&\times \exp\left(-\frac{1}{2}\sum_{\ell=1}^{g}\mathsf{1}_{z=\ell}\left\{\mu_\ell^T\Sigma^{-1}\mu_\ell - 2\log\alpha_\ell\right\}\right)
\end{aligned}$$

Hence, $F(\theta)$ is of the form

$$-\frac{1}{n}\sum_{i=1}^{n}\log\int_{\mathsf{Z}} h_i(z)\ \exp(\langle s_i(z), \phi(\theta)\rangle)\,\mu(\mathrm{d}z) + \mathsf{R}(\theta)$$

by setting: μ the counting measure on $[g]^\star$, $\mathsf{Z} = [g]^\star$, $h_i(z) = 1$, and

$$\mathsf{R}(\theta) \stackrel{\text{def}}{=} \frac{1}{2}\mathrm{Trace}\left(\Sigma^{-1}\frac{1}{n}\sum_{i=1}^{n}Y_iY_i^T\right) - \frac{1}{2}\log\det(\Sigma^{-1}),$$

$$\phi(\theta) \stackrel{\text{def}}{=} \begin{bmatrix} \log\alpha_1 - 0.5\mu_1^T\Sigma^{-1}\mu_1 \\ \cdots \\ \log\alpha_g - 0.5\mu_g^T\Sigma^{-1}\mu_g \\ \Sigma^{-1}\mu_1 \\ \cdots \\ \Sigma^{-1}\mu_g \end{bmatrix}, \quad s_i(z) \stackrel{\text{def}}{=} \begin{bmatrix} \mathsf{1}_{z=1} \\ \cdots \\ \mathsf{1}_{z=g} \\ \mathsf{1}_{z=1}Y_i \\ \cdots \\ \mathsf{1}_{z=g}Y_i \end{bmatrix}.$$

Note that $\phi(\theta) \in \mathbb{R}^{g+pg}$ and $s_i(z) \in \mathbb{R}^{g+pg}$.

From these expressions, for any $i \in [n]^\star$ and $\theta \in \Theta$, the distribution $z \mapsto p_i(z;\theta)$ is the distribution on $[g]^\star$ given by

$$p_i(\ell;\theta) = \frac{\alpha_\ell \exp\left(-\frac{1}{2}(Y_i-\mu_\ell)^T\Sigma^{-1}(Y_i-\mu_\ell)\right)}{\sum_{u=1}^{g}\alpha_u\exp\left(-\frac{1}{2}(Y_i-\mu_u)^T\Sigma^{-1}(Y_i-\mu_u)\right)}$$

for any $\ell \in [g]^\star$. Therefore

$$\bar{s}_i(\theta) = \begin{bmatrix} p_i(1;\theta) \\ \cdots \\ p_i(g;\theta) \\ p_i(1;\theta)Y_i \\ \cdots \\ p_i(g;\theta)Y_i \end{bmatrix}.$$

For a vector $s \in \mathbb{R}^{g+pg}$, we write

$$s = \begin{bmatrix} s^{(1),1} \\ \cdots \\ s^{(1),g} \\ s^{(2),1} \\ \cdots \\ s^{(2),g} \end{bmatrix},$$

where $s^{(1),\ell} \in \mathbb{R}$ and $s^{(2),\ell} \in \mathbb{R}^p$; with these notations, $\mathsf{T}(s) = (\alpha_1, \cdots, \alpha_g, \mu_1, \cdots, \mu_g, \Sigma)$ is defined by

$$\begin{aligned}
\alpha_\ell &\stackrel{\text{def}}{=} \frac{s^{(1),\ell}}{\sum_{u=1}^{g} s^{(1),u}} \\
\mu_\ell &\stackrel{\text{def}}{=} \frac{s^{(2),\ell}}{s^{(1),\ell}} \\
\Sigma &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}Y_iY_i^T - \sum_{\ell=1}^{g} s^{(1),\ell}\mu_\ell\mu_\ell^T.
\end{aligned}$$

2.2. Additional plots

In the numerical applications, $p = 20$, $g = 12$ and $n = 6 \times 10^4$.

Figure 1 displays the number of conditional expectation evaluations per epoch (top) or cumulated vs the number of epoch (bottom). For `quad-geom`, the total number of conditional expectations (CE) evaluations is

- $\lfloor(\min(n, \max(n/100, 6t^2)))\rfloor$ for the computation of S_t
- and the 2b multiplied by the mean number of ξ_t.

For `full-geom`, the total number of CE evaluations is

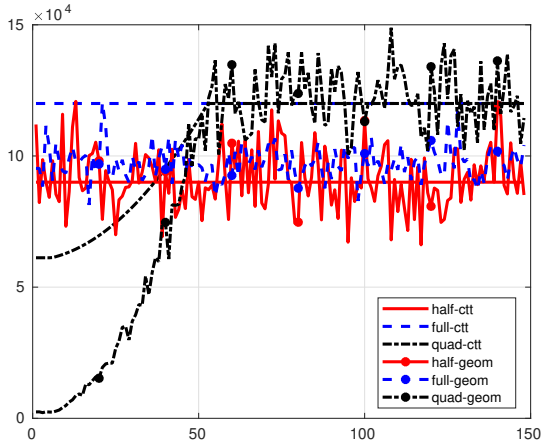
- n for the computation of S_t
- and the $2x$b multiplied by the mean number of ξ_t.

For `full-ctt`,

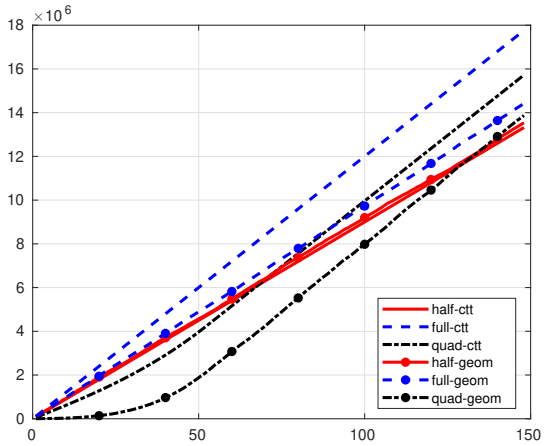
- n for the computation of S_t
- and the $2x\mathsf{b}$ multiplied by k_{in} which is constant over the inner loops.

For `quad-ctt`,

- $\lfloor(\min(n, \max(n/100, 6t^2)))\rfloor$ for the computation of S_t
- and the $2x\mathsf{b}$ multiplied by the mean number of ξ_t.



(a) Per epoch

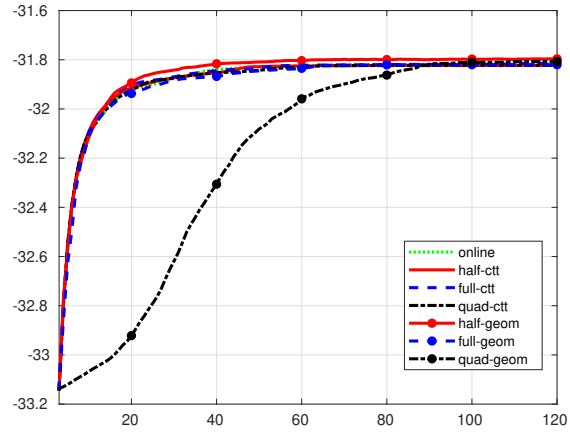


(b) cumulated

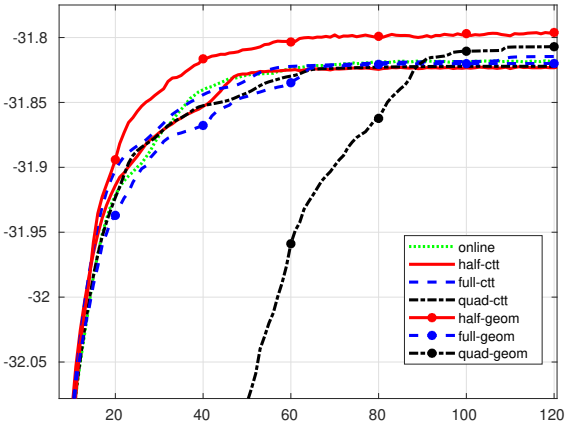
Fig. 1. Number of CE evaluations: per epochs (top) and cumulated (bottom).

Figure 2 displays the mean value of the normalized log-likelihood $n^{-1}\sum_{i=1}^n \mathcal{L}_i \circ \mathsf{T}(\widehat{S}_{t,\Xi_t})$ vs the number of epochs t; this mean value is computed over 30 independent runs of the algorithms. Figure 3 displays the same analysis as in Figure 2 for two different strategies for the number of inner loops ξ_t (on the top, ξ_t is constant and equal to k_{in}; on the bottom, ξ_t is a geometric distribution with expectation k_{in}), and different strategies for the initialization $\mathcal{E}_t$.

Figure 4 displays the quantiles 0.25, 0.5 and 0.75 of the distribution of $\|h(\widehat{S}_{t,\Xi_t})\|^2$; the quantiles are estimated over 30 independent



(a) vs epoch 1 to 120



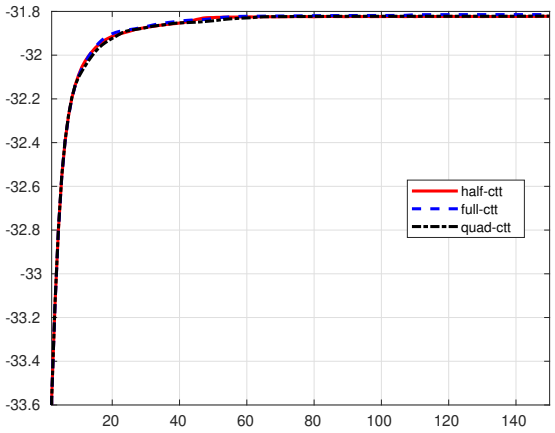
(b) the first epochs are discarded

Fig. 2. The mean value, computed over 30 independent runs of the algorithms, of the normalized log-likelihood $n^{-1}\sum_{i=1}^n \mathcal{L}_i$. The plot shows its value vs the number of epochs.

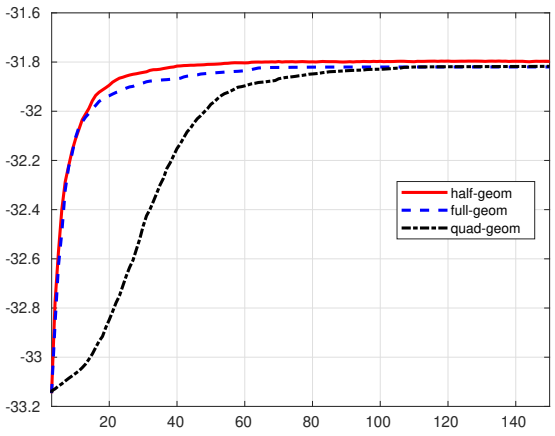
runs of the algorithm. Here the case ξ_t is constant is considered, with different strategies for the initialization $\mathcal{E}_t$.

Figure 5 displays the same analysis as in Figure 4, except that ξ_t is a geometric random variable with expectation k_{in}.

Figure 6 display Figure 4 and Figure 5 on the same plots. The quantiles of the distribution of $\|h(\widehat{S}_t)\|^2$ vs the number of epochs when $\{\widehat{S}_t, t \geq 0\}$ is obtained by `online-EM`.

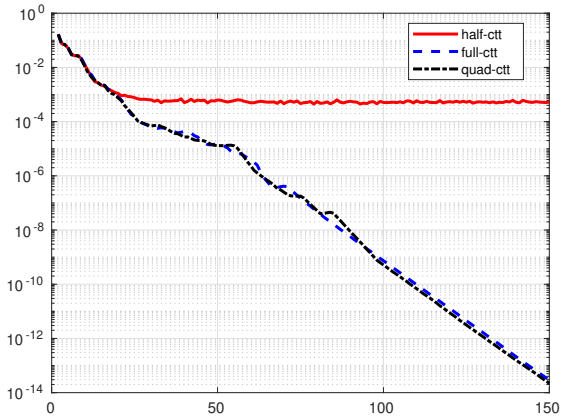


(a) Case of SPIDER-EM (fixed number of inner loop $\xi_t = k_{\text{in}}$)

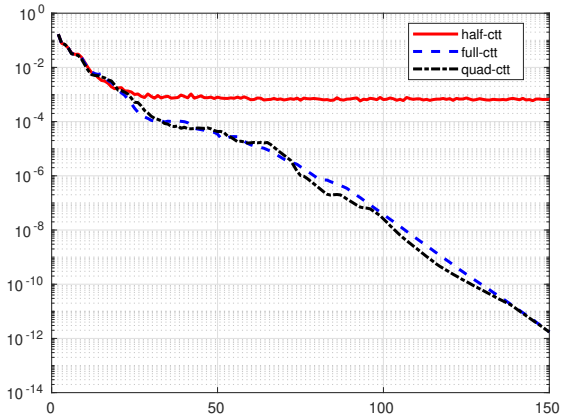


(b) Case of Geom-SPIDER-EM

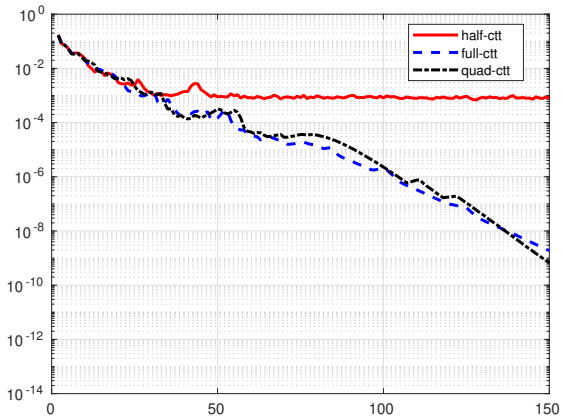
Fig. 3. The mean value, computed over 30 independent runs of the algorithms, of the normalized log-likelihood $n^{-1}\sum_{i=1}^{n}\mathcal{L}_i$. The plot shows the mean value vs the number of epochs, starting from epoch #2



(a) Quantile 0.25

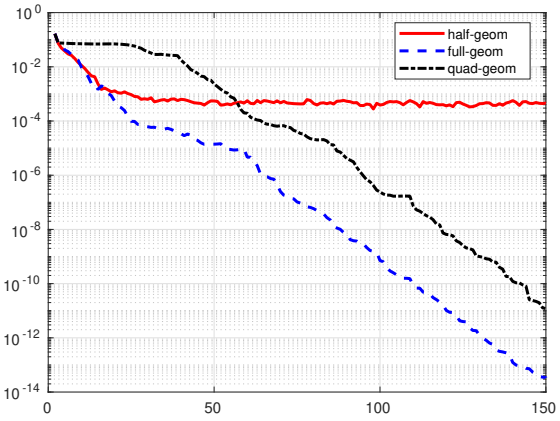


(b) Quantile 0.50

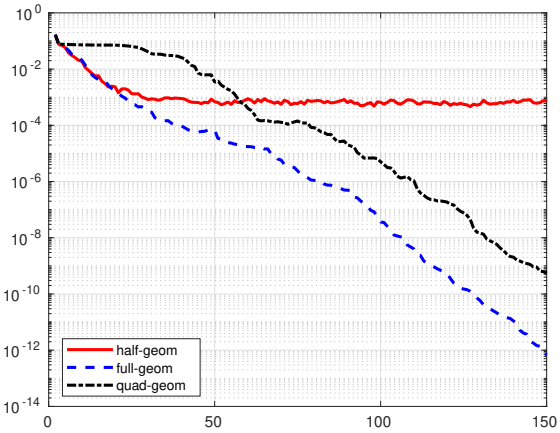


(c) Quantile 0.75

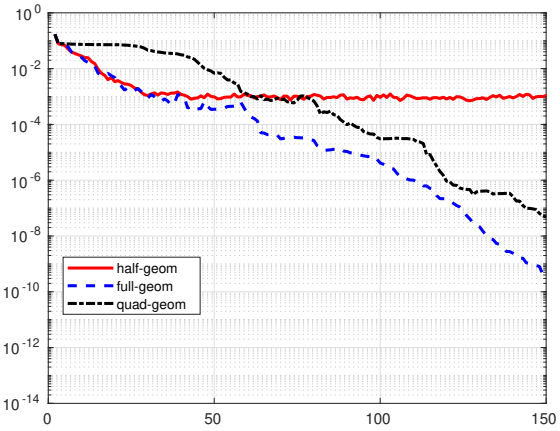
Fig. 4. Case SPIDER-EM. Quantiles of the distribution of $\|h(\widehat{S}_{t,\Xi_t})\|^2$, estimated over 30 independent runs, vs the number of epochs t.



(a) Quantile 0.25

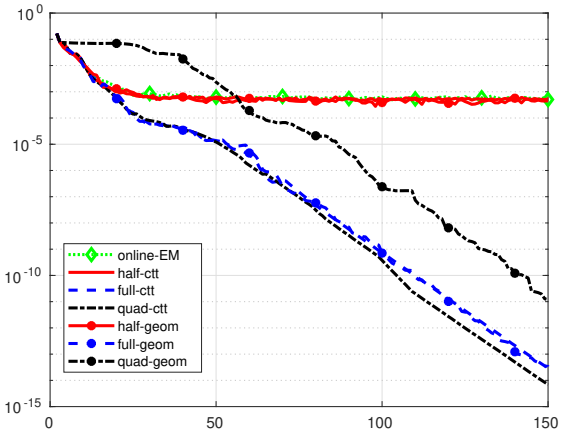


(b) Quantile 0.50

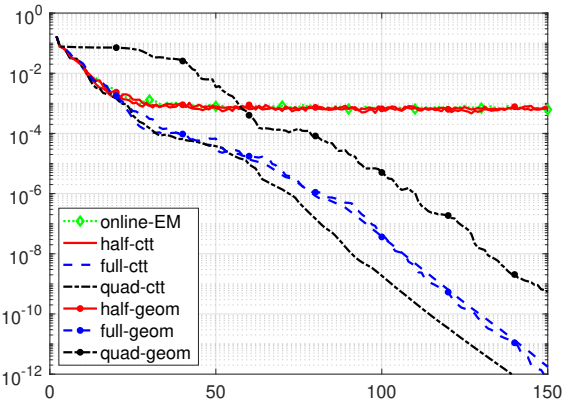


(c) Quantile 0.75

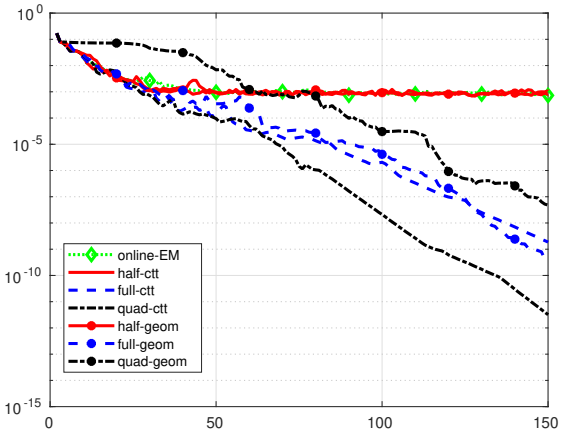
Fig. 5. Case Geom-SPIDER-EM. Quantiles of the distribution of $\|h(\widehat{S}_{t,\Xi_t})\|^2$, estimated over 30 independent runs, vs the number of epochs t



(a) Quantile 0.25



(b) Quantile 0.50



(c) Quantile 0.75

Fig. 6. Comparison of Online-EM, SPIDER-EM and Geom-SPIDER-EM. Quantiles of the distribution of $\|h(\widehat{S}_{t,\Xi_t}\|^2$, estimated over 30 independent runs, vs the number of epochs t.

3. REFERENCES

[1] G. Fort, E. Moulines, and H.T. Wai, "A Stochastic Path-Integrated Differential EstimatoR Expectation Maximization Algorithm," in *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc., 2020.